ESTIMATION OF A FUNCTION WITH DISCONTINUITIES VIA LOCAL POLYNOMIAL FIT WITH AN ADAPTATIVE WINDOW CHOICE

SPOKOINY, V.G.

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin and
Institute for Information Transmission Problems,
Bolshoi Karetnyi 19, Moscow 101447, Russia

1991 *Mathematics Subject Classification.* 62G07; Secondary 62G20.

Key words and phrases. change-point, local polynomial fit local structure, nonparametric regression, pointwise adaptive estimation .

The author thanks O. Lepski, A. Juditski and M. Neumann for helpful remarks and discussion.

Abstract. New method of adaptive estimation of a regression function is proposed. The resulting estimator achieves near optimal rate of estimation in the classical sense of mean integrated squared error. At the same time, the estimator is shown to be very sensitive to discontinuities or change-points of the underlying function f or its derivatives. For instance, in the case of a jump of a regression function, beyond the interval of length (in order) $n^{-1}\log n$ around change-points the quality of estimation is essentially the same as if the location of this jump were known. The method is fully adaptive and no assumptions are imposed on the design, number and size of jumps. The results are formulated in a non-asymptotic way and can be therefore applied for an arbitrary sample size.

1. Introduction

The change-point analysis which includes sudden, localized changes typically occurring in economics, medicine and the physical sciences has recently found increasing interest, see Müller (1992) for some examples and discussion of the problem. There are several aspects of change-point problems addressed in the literature which rely on

- type of the model (regression models, distribution or spectral density models etc.),
- type of observations (discrete or continuous),
- assumptions about discontinuities incorporated in the model:
  - type of discontinuities (jumps, change-points, cusps)
  - one or more change-points;
  - the number and sizes (magnitude) of change-points are known or not;
- assumption on the model function (parametric or nonparametric).

We restrict ourselves to change-points in the regression model with nonparametrically described regression function. A comparison of parametric and nonparametric approaches lies beyond the scope of this paper. We note only that the nonparametric approach provides with much more flexibility without strong influence of the quality of estimation, for more discussion see Müller (1992).

The choice of regression model is motivated by statistical practice, especially in econometrics. A discussion of related problems for other models (for instance, distribution density model) allowing change-points is a subject of another paper.

In the regression nonparametric analysis of function with change-points, one may highlight two different directions . The first approach deals with a generally smooth curve allowing a finite number of change-points. Further one may focus either on estimation of locations and magnitudes of jumps, as in Korostelev (1987), Yin (1988), Wang (1995), or on estimating the function itself. In the last case, some pilot near optimal estimates of locations of change-points are still required as a technical step in the estimation procedure. When all the locations of change-points have been estimated, one may estimate the function separately on each interval between every neighbor points, see Müller (1992), Wu and Chu (1993), Oudshoorn (1995). The most remarkable fact here, due to Korostelev (1987), is that the location of a single jump of a given magnitude can be estimated with the rate n^{-1} where n is the number of observations. This result can be generalized on the case with unknown jump size and the case of jump of some derivative of the function f, Müller (1992), and even on the case with a finite unknown number of change-points of different order, Yin (1988), Oudshoorn (1995). As a price for such kind of adaptation, the rate of estimating the locations of jumps is worse by some logarithmic factor. The location of a jump of kth derivatives can be estimated with the rate about $n^{-1/(2k+1)}$ multiplied again by some log-factor. However, this rate is still much better than in estimating the corresponding derivative of the regression function and such sort of procedures allows to perform asymptotically optimal estimation of a regression function with change-points, Oudshoorn (1995).

Another approach to this problem is connected with the concept of spatial adaptive estimation. The problem of adaptive and spatially adaptive nonparametric estimation is now well developed, see e.g. Nemirovski (1985), Donoho et al (1994), Lepski, Mammen and Spokoiny (1994), Delyon and Juditski (1994), Goldenshluger and Nemirovski (1994), Lepski and Spokoiny (1995) among others. A variety of different adaptive methods can be now applied to estimation of a function with inhomogeneous smoothness characteristics: non-linear wavelet procedure, kernel estimation with a variable bandwidth, local polynomials with a variable window etc. For all such methods it is shown near optimality in the minimax sense for mean integrated squared errors.

In the context of spatially adaptive nonparametric estimation, change-points or, more generally, cusps in the curve can be viewed as a sort of inhomogeneous behavior of the estimated function. One may therefore apply the same procedure (for instance nonlinear wavelet estimation) and the analysis is focusing on the quality of estimation when change-points are incorporated in the model. Under this approach, the main intention is to estimate the regression function (not locations of change-points). It is shown in Hall and Patil (1995), Hall, Kerkyacharian and Picard (1996) that the wavelet-based estimators provide the same rate of estimation even if a growing number of jumps is allowed.

The important advantage of this approach is that the estimating procedure is universal and it is not specified to change-point analysis. On the other side, this approach delivers very poor qualitative information about presence, number and locations of change-points. Moreover, the criteria based on mean integrated errors are not very sensitive to local quality of estimation: having obtained the optimal rate in global estimation, one gets relatively poor quality of estimation in small vicinities of change-points. The reason for that can be explained on the example of kernel estimation with a variable bandwidth. It was shown in Lepski, Mammen and Spokoiny (1994) that there is a procedure of such kind which provides the optimal rate of estimation over a wide range of Besov classes containing particularly functions with change-points. However, this procedure applies everywhere a symmetric kernel and the resulting estimator has a "boundary (or Gibbs) effect" in a vicinity of radius h around each change-point, h being the applied bandwidth. At the same time, imagine that the locations of change-points were known. Then it would be obviously much better to apply one-side kernels near change-points.

The aim of the present paper is to propose a method which simultaneously adapts to inhomogeneous smoothness of the estimated curve and which is sensitive to discontinuities of the curve or its derivatives. We develop the pointwise adaptive approach by selecting not only the applied bandwidth but also the geometry of the window: we search for a maximal window containing the point of estimation in which the function f is "smooth". (This can be understood in the sense that it is well approximated by polynomials.) Such a procedure selects automatically between windows without change-points.

We consider the regression model

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where $X_i \in R^1$, $i = 1, \ldots, n$, are given design points and ξ_i are individual independent random errors. Below we will suppose that ξ_i, $i = 1, \ldots, n$, are i.i.d. $\mathcal{N}(0, \sigma^2)$ with a given noise level σ. This assumption can be easily relaxed in a usual way.

The idea of the proposed method is quite simple and natural. We assume that the function f is well approximated by a polynomial $P_\theta(\cdot - x_0)$ in some neighborhood U of the point of interest x_0, where θ is the vector of coefficients of this polynomial. Basing on this assumption, for each feasible interval U containing x_0, one can construct an estimator $\hat{\theta}$ of θ by the observations $\{Y_i, X_i : X_i \in U\}$ and then calculate the residuals $\varepsilon_i = Y_i - P_{\hat{\theta}}(X_i - x_0)$. Next we test the hypothesis that the residuals $\varepsilon_i = \varepsilon_i(X_i)$ can be treated on the interval U as a pure noise. Finally the procedure selects the maximal (in length) interval for which this hypothesis is not rejected. We show that this method provides both spatial adaptive estimation in the sense of mean integrated squared losses and high sensitivity to change-points of f.

The benefit of this approach is that it is very general in nature and it is not specified to estimation of a regular function with change-points. One may therefore expect that this method can be extended on the case of multi-dimensional regression or applied to image denoising where the quality of estimation near the boundary of images is of special importance, see Korostelev and Tsybakov (1994). One more important feature of the proposed pointwise approach is that it allows to proceed with an arbitrary design. We do not need to assume random or regular (for instance equidistant) design. With it, the resulting quality of estimation depends on the local design properties near the point x_0.

The paper is organized as follows. In the next section we present the procedure, Section 3 contains the results describing the quality of this procedure. In Section 4 we specify the general results to the case of the equidistant design. We show in particular that the locations of jumps can be estimated with the rate $n^{-1} \log n$ and that this rate cannot be improved if more than one jump is allowed.

2. Estimation Procedure

Let data Y_i, X_i, $i = 1, \ldots, n$ obeys the model (1.1). We will estimate $f(x_0)$ for a given x_0.

First we describe the family $\mathcal{U}$ of intervals containing x_0. This family can be introduced in different ways. One possible choice is to consider all intervals with the edges at design points,

$$\mathcal{U} = \{[X_i, X_{i'}] : X_i \le x_0 \le X_{i'}\}. \tag{2.1}$$

This choice is theoretically possible and it allows to make very precise estimation but it leads to a serious computational effort. One may decrease the cardinality of $\mathcal{U}$ and hence the computational difficulties by selecting two sets of points $\mathcal{A}_l = \{a_l : a_l \le x_0\}$ and $\mathcal{A}_r = \{a_r : a_r \ge x_0\}$ and by setting

$$\mathcal{U} = \{U = [a_l, a_r] : a_l \in \mathcal{A}_l,\ a_r \in \mathcal{A}_r,\ N_U \ge m\}. \tag{2.2}$$

The sets $\mathcal{A}_l$ and $\mathcal{A}_r$ should be "dense" enough near x_0 to provide the desirable quality of estimation. For instance, for the case of a regular or random design with a positive density, one may take $\mathcal{A}_l$ and $\mathcal{A}_r$ in the form of geometrical grids, $\mathcal{A}_l = \{x_0 - ab^{-k}, 0 \le k \le \log n\}$ and similarly for $\mathcal{A}_r$. Here $a > 0$ and $b > 1$. In this case, the set $\mathcal{U}$ has cardinality $\log^2 n$. For an arbitrary design, one can define the sets $\mathcal{A}_l$ and $\mathcal{A}_r$ in a similar way providing that each U from $\mathcal{U}$ contains at least m design points X_i.

Given $U \in \mathcal{U}$, set N_U for the number of the points X_i falling in U,

$$N_U = \#\{X_i : X_i \in U\}.$$

We will suppose that $N_U \ge m$ for each $U \in \mathcal{U}$.

Now we construct by data $\{Y_i, X_i : X_i \in U\}$ a polynomial P of degree $m-1$ which approximates the underlying function f on U. For this we apply the standard least squared method. We consider first the general case with an arbitrary integer m and then specify the procedure for $m = 2$ which seems to be reasonable for practical applications.

Let $m \ge 1$ be given. By θ we will denote a column-vector in R^m, $\theta = (\theta_0, \ldots, \theta_{m-1})^T$ and by $P_\theta(z)$ the polynomial with the coefficients θ, $P_\theta(z) = \theta_0 + \theta_1 z + \ldots + \theta_{m-1} z^{m-1}$. Define $\hat{\theta}_U$ by the least squared method

$$\hat{\theta}_U := \operatorname*{arginf}_{\theta} \sum_U (Y_i - P_\theta(X_i - x_0))^2.$$

For an explicit representation of $\hat{\theta}_U$, it is useful to introduce matrix notation. Let Σ_U be the $m \times N_U$-matrix with elements $s_{k,i} = (X_i - x_0)^k$, $k = 0, 1, \ldots, m-1$, and let Y_U be the N_U-column vector with elements Y_i where only indices i with $X_i \in U$ are considered. Then the vector $\hat{\theta}_U$ satisfies the equation

$$\Sigma_U \Sigma_U^T \hat{\theta}_U = \Sigma_U Y_U. \tag{2.3}$$

If the matrix $D_U = N_U^{-1} \Sigma_U \Sigma_U^T$ is non-singular that is $\det D_U \ne 0$, then $\hat{\theta}_U$ can be defined by

$$\hat{\theta}_U = (\Sigma_U \Sigma_U^T)^{-1} \Sigma_U Y_U, \tag{2.4}$$

Otherwise one can use the same representation, understanding $(\Sigma_U \Sigma_U^T)^{-1}$ as pseudo-inverse matrix.

The vector $\hat{\theta}_U$ provides with the non-parametric estimators of the function f and its derivatives at x_0. Namely, one can use for estimation of f the value of the approximating polynomial $P_{\hat{\theta}_U}$ and its derivatives at x_0. Thus, $k!\hat{\theta}_{U,k}$ is the estimator of $f^{(k)}(x_0)$. Particularly, $\hat{f}_U(x_0) = \hat{\theta}_{U,0}$ is the estimator of $f(x_0)$.

The residuals $\varepsilon_{U,i}$ at points $X_i \in U$ are defined by $Y_i - P_{\hat{\theta}_U}(X_i - x_0)$, that is

$$\varepsilon_{U,i} = Y_i - \hat{\theta}_{U,0} - \hat{\theta}_{U,1}(X_i - x_0) - \ldots - \hat{\theta}_{U,m-1}(X_i - x_0)^{m-1}.$$

Using matrix notation, one gets

$$\varepsilon_U = Y_U - \Sigma_U^T \hat{\theta}_U = Y_U - \Sigma_U^T (\Sigma_U \Sigma_U^T)^{-1} \Sigma_U Y_U = Y_U - \Pi_U Y_U. \tag{2.5}$$

Note that $\Pi_U = \Sigma_U^T (\Sigma_U \Sigma_U^T)^{-1} \Sigma_U$ is projector from the space R^{N_U} on the linear subspace generated by polynomials of degree $m-1$ with knots $X_i \in U$.

Our adaptation method is based on the analysis of the residuals $\varepsilon_{U,i}$. We introduce another family $\mathcal{V}(U)$ of intervals V, each of them is a subinterval of U. As above for the family $\mathcal{U}$, we require that $N_V := \#\{X_i \in V\} \geq m$ for all $V \in \mathcal{V}(U)$. Also we require that $V = U \cap U' \in \mathcal{V}(U)$ for each $U' \in \mathcal{U}$.

A reasonable way to define this family is as follows

$$\mathcal{V}(U) = \{V = U \backslash U' \text{ or } V = U \cap U' : U' \in \mathcal{U},\ N_V \geq m\}.$$

If the set $\mathcal{U}$ is of the form (2.2), then one has obviously

$$\mathcal{V}(U) = \{V = [a_-, a_+] : a_-, a_+ \in \mathcal{A}_l \cup \mathcal{A}_r,\ V \subseteq U,\ N_V \geq m\}. \tag{2.6}$$

Below we need in some upper estimate of the cardinality of $\mathcal{V}(U)$ in the form

$$\#\mathcal{V}(U) \leq N_U^\alpha \tag{2.7}$$

with some $\alpha > 0$. In the case of "maximal" $\mathcal{U}$ from (2.1), and with $\mathcal{V}(U)$ from (2.6), one can easily check (2.7) for $\alpha = 4$.

For each $V \in \mathcal{V}(U)$ and for every $k = 0, 1, \dots, m-1$, set

$$T_{U,V,k} = \frac{1}{\sigma \sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k \varepsilon_{U,i},$$

where

$$d_{V,k} = \frac{1}{N_V} \sum_V (X_i - x_0)^k, \qquad k = 0, 1, \dots, 2m \tag{2.8}$$

and $\sum_V$ means summation over the index set $\{i : X_i \in V\}$.

Define now

$$\varrho_{U,V} = \mathbf{1}\left(\max_{0 \leq k \leq m-1} |T_{U,V,k}| > t\sqrt{\log N_U} \right)$$

where

$$t = (2 + \sqrt{m})\sqrt{2(\alpha + p)}.$$

The parameter p has meaning of the norm in which we measure loss of estimation. Typically one sets $p = 2$.

We say that U is rejected if $\varrho_{U,V} = 1$ at least for one $V \in \mathcal{V}(U)$ i.e. if $\varrho_U = 1$ where

$$\varrho_U = \sup_{V \in \mathcal{V}(U)} \varrho_{U,V} = \mathbf{1}\left(\sup_{V \in \mathcal{V}(U)} \max_{0 \leq k \leq m-1} |T_{U,V,k}| > t\sqrt{\log N_U} \right).$$

The adaptive procedure selects among all non-rejected U from $\mathcal{U}$ such one which maximizes N_U,

$$U^* = \operatorname*{argmax}_{U \in \mathcal{U}} \{N_U : \varrho_{U,V} = 0 \text{ for all } V \in \mathcal{V}(U)\} \tag{2.9}$$

and

$$\hat{f}(x_0) = \hat{f}_{U^*}(x_0) = \hat{\theta}_{U^*,0}. \tag{2.10}$$

For technical reason, we need to bound the considered class of functions. Namely we suppose that the function f is bounded in absolute value by some known constant f_0. Accordingly we truncate the estimate $\hat{f}(x_0)$ from (2.10), i.e. we apply the estimate $-f_0 \vee \hat{f}(x_0) \wedge f_0$.

2.1. The case with $m = 2$

Below we specify the above procedure for the case of locally linear approximation when $m = 2$.

For a fixed $U \in \mathcal{U}$, one has

$$\begin{aligned} d_{U,0} &= 1, \\ d_{U,1} &= \frac{1}{N_U} \sum_U (X_i - x_0), \\ d_{U,2} &= \frac{1}{N_U} \sum_U (X_i - x_0)^2, \end{aligned}$$

and also

$$\begin{aligned} B_{U,0} &= \frac{1}{N_U} \sum_U Y_i, \\ B_{U,1} &= \frac{1}{N_U} \sum_U (X_i - x_0) Y_i. \end{aligned}$$

Now

$$\begin{aligned} \hat{\theta}_{U,0} &= \frac{d_{U,2} B_{U,0} - d_{U,1} B_{U,1}}{d_{U,2} - d_{U,1}^2}, \\ \hat{\theta}_{U,1} &= \frac{B_{U,1} - d_{U,1} B_{U,0}}{d_{U,2} - d_{U,1}^2}, \\ \varepsilon_{U,i} &= Y_i - \hat{\theta}_{U,0} - \hat{\theta}_{U,1}(X_i - x_0). \end{aligned}$$

Next, for every $V \in \mathcal{V}(U)$, we have

$$\begin{aligned} T_{U,V,0} &= \frac{1}{\sigma \sqrt{N_V}} \sum_V \varepsilon_i, \\ T_{U,V,1} &= \frac{1}{\sigma \sqrt{d_{V,2} N_V}} \sum_V (X_i - x_0) \varepsilon_i, \\ \varrho_U &= \mathbf{1} \left(\sup_{V \in \mathcal{V}(U)} \max_{k=0,1} |T_{U,V,k}| > 3.5 \sqrt{2(\alpha + p) \log N_U} \right). \end{aligned}$$

The adaptive estimator $\hat{f}(x_0)$ is defined now by (2.9) and (2.10).

3. Main results

In this section we describe some properties of the proposed estimation procedure. We distinguish between two extreme cases: the function f is regular in a usual sense near the point of interest x_0 or this function has a jump in the nearest vicinity of this point.

To formulate the results, we introduce an important characteristic of the function f which describes the accuracy of approximation of f by polynomials. Given $U \in \mathcal{U}$, define $\Delta_U(f)$ by

$$\Delta_U(f) = \inf_{P \in \mathcal{P}_m} \sup_{x \in U} |f(x) - P(x - x_0)|$$

where $\mathcal{P}_m$ is the set of all polynomials of degree $m-1$. Obviously $\Delta_{U'}(f) \le \Delta_U(f)$ if $U' \subset U$.

The first results claims that if f is "smooth" on U in the sense that $\Delta_U(f)$ is small enough then the procedure rejects U with very small probability.

Proposition 3.1. *Let $U \in \mathcal{U}$ be such that*

$$\Delta_U(f) \le C_1(\sigma^2 N_U^{-1} \log N_U)^{1/2} \tag{3.1}$$

where

$$C_1 = \sqrt{2(\alpha + p)}$$

Then

$$\mathbf{P}_f(\varrho_U = 1) \le m N_U^{-p}.$$

Basing on this result, we denote by $\mathcal{U}^+$ the subset of $\mathcal{U}$ whose elements U obey (3.1),

$$\mathcal{U}^+ = \{U \in \mathcal{U} : \Delta_U^2(f) \le 2\sigma^2(\alpha + p) N_U^{-1} \log N_U\}. \tag{3.2}$$

Proof. Using the model equation (1.1), rewrite the vector of residuals ε_U in the form

$$\varepsilon_U = f_U - \Pi_U f_U + \xi_U - \Pi_U \xi_U = f_U - \Pi_U f_U + \xi_U - \zeta_U,$$

see (2.5). Here f_U means the vector with elements $f(X_i)$, $X_i \in U$, and $\zeta_U = \Pi_U \xi_U$. The "test" statistic $T_{U,V,k}$ can be represented now in the form

$$\begin{aligned} T_{U,V,k} &= \frac{1}{\sigma\sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k (f(X_i) - \Pi_U f(X_i)) + \\ &\quad \frac{1}{\sigma\sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k \xi_i - \frac{1}{\sigma\sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k \zeta_U(X_i) \\ &= S_1 + S_2 + S_3. \end{aligned} \tag{3.3}$$

We analyze each sum in this expression separately starting from the first one.

By definition of $\Delta_U(f)$, there exists for each $\gamma > 0$ a polynomial $P \in \mathcal{P}_m$ such that $\sum_U |f(X_i) - P(X_i - x_0)|^2 \le N_U \Delta_U^2(f) + \gamma$. To simplify the exposition, we will suppose that this inequality holds with $\gamma = 0$. Since Π_U is the projector on the space generated by polynomials of degree $m-1$, then one has $\Pi_U P = P$ and hence

$$\|f-\Pi_U f\|_U^2 = \|f-P-\Pi_U(f-P)\|_U^2 \le \|f-P\|_U^2 \le N_U \Delta_U^2(f)$$

where $\|f\|_U^2 = \sum_U f^2(X_i)$. Now we get using Cauchy-Schwarz inequality and the condition (3.1)

$$\begin{aligned}
S_1 &= \frac{1}{\sigma\sqrt{d_{V,2k}N_V}} \sum_V (X_i-x_0)^k (f(X_i)-\Pi_U f(X_i)) \\
&\le \left[\frac{1}{\sigma^2 d_{V,2k}N_V}\sum_V (X_i-x_0)^{2k}\right]^{1/2} \left[\sum_V (f(X_i)-\Pi_U f(X_i))^2\right]^{1/2} \\
&\le \sigma^{-1}\|f-\Pi_U f\|_V \le \sigma^{-1}\|f-\Pi_U f\|_U \le \sigma^{-1}\sqrt{N_U}\Delta_U(f) \\
&\le \sqrt{2(\alpha+p)\log N_U}. \qquad (3.4)
\end{aligned}$$

Next, since the random errors ξ_i are Gaussian zero mean, the same is true for the sum S_2 in (3.3). Moreover, using independence of ξ_i's,

$$\mathbf{E}S_2^2 = \frac{1}{\sigma^2 d_{V,2k}N_V}\sum_V (X_i-x_0)^{2k}\mathbf{E}\xi_i^2 = 1 \qquad (3.5)$$

and hence S_2 is standard Gaussian.

It remains to estimate S_3. The vector $\zeta_U = \Pi_U \xi_U$ is Gaussian as the linear transform of the Gaussian vector ξ_U. Obviously $\mathbf{E}\zeta_U = 0$. Moreover,

$$\begin{aligned}
\mathbf{E}\zeta_U\zeta_U^T &= \Sigma_U^T(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U \mathbf{E}\xi_U\xi_U^T \Sigma_U^T(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U \\
&= \sigma^2\Sigma_U^T(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U\Sigma_U^T(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U = \\
&= \sigma^2\Sigma_U^T(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U.
\end{aligned}$$

Here we have used that $\mathbf{E}\xi_i\xi_j = \sigma^2\delta_{i,j}$. This implies

$$\begin{aligned}
\sum_U \mathbf{E}\zeta_U^2(X_i) &= \operatorname{tr}\mathbf{E}\zeta_U\zeta_U^T \\
&= \sigma^2 \operatorname{tr}\Sigma_U^T(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U \\
&= \sigma^2 \operatorname{tr}(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U\Sigma_U^T \\
&\le \sigma^2 \operatorname{tr} I_m = \sigma^2 m
\end{aligned}$$

where $\operatorname{tr}A$ is set for the trace of matrix A and I_m means the unit $m\times m$-matrix.

Now, using again the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\mathbf{E}S_3^2 &= \frac{1}{\sigma^2 d_{V,2k}N_V}\mathbf{E}\left[\sum_V (X_i-x_0)^k\zeta_U(X_i)\right]^2 \\
&\le \left[\frac{1}{\sigma^2 d_{V,2k}N_V}\sum_V (X_i-x_0)^{2k}\right]\left[\sum_V \mathbf{E}\zeta_U^2(X_i)\right] \\
&\le \sigma^{-2}\sum_U \mathbf{E}\zeta_U^2(X_i) \le m. \qquad (3.6)
\end{aligned}$$

The sum of two Gaussian variables S_2 and S_3 is also Gaussian with zero mean and along with (3.5), (3.6)

$$\begin{aligned}\mathbf{E}(S_2+S_3)^2 &= \mathbf{E}S_2^2+\mathbf{E}S_3^2+2\mathbf{E}S_2S_3\\ &\le \mathbf{E}S_2^2+\mathbf{E}S_3^2+2(\mathbf{E}S_2^2\mathbf{E}S_3^2)^{1/2}\\ &\le (1+\sqrt{m})^2.\end{aligned}$$

Summing up (3.4) through (3.6), we get

$$\begin{aligned}&\mathbf{P}_f\left(|T_{U,V,k}|>(2+\sqrt{m})\sqrt{2(\alpha+p)\log N_U}\right)\\ &\quad\le \mathbf{P}\left(|S_2+S_3|>(1+\sqrt{m})\sqrt{2(\alpha+p)\log N_U}\right)\\ &\quad\le 2\left(1-\Phi\left(\sqrt{2(\alpha+p)\log N_U}\right)\right)\\ &\quad\le \exp\{-(\alpha+p)\log N_U\}=N_U^{-(\alpha+p)}.\end{aligned}$$

Here Φ means the Laplace distribution and we have used that $1-\Phi(z)\le\exp(-z^2/2)$ for $z>0$. This estimate and the condition (2.7) allow to bound the probability of rejecting U in the following way

$$\begin{aligned}\mathbf{P}_f(\varrho_U=1) &\le \sum_{V\in\mathcal{V}(U)}\sum_{k=0}^{m-1}\mathbf{P}_f\left(|T_{U,V,k}|>(2+\sqrt{m})\sqrt{2(\alpha+p)\log N_U}\right)\\ &\le m\,\#\mathcal{V}(U)\,N_U^{-(\alpha+p)}\le mN_U^{-p}\end{aligned}$$

as required. □

An interesting feature of the above result is that no assumptions were made about the design on U. For the next statement, as usual for the local polynomial estimation, we introduce some condition on the design. Given $U\in\mathcal{U}$, denote by G_U the $m\times m$-matrix with elements $g_{U,k,k'}=d_{U,k+k'}/\sqrt{d_{U,2k}d_{U,2k'}}$, $k,k'=0,1,\ldots,m-1$, see (2.8). It is convenient to use the following matrix notation. Let Λ_U be the diagonal matrix with diagonal elements $d_{U,2k}^{-1/2}$,

$$\Lambda_U=\mathrm{diag}(1,d_{U,2}^{-1/2},\ldots,d_{U,2m-2}^{-1/2})$$

Then

$$G_U=\Lambda_U D_U\Lambda_U. \tag{3.7}$$

Set also $|\det G_U|$ for the absolute value of the determinant of G_U. It is easy to see that $|\det G_V|\le|\det G_U|$ for $V\subset U$.

Our condition on the design means that $|\det G_U|$ is bounded away from zero. For this is typically enough to have m design points in general position inside the interval U.

Proposition 3.2. *Let $U\in\mathcal{U}$, $V\in\mathcal{V}(U)$ and let $\varrho_{U,V}=0$. If $|\det G_V|>0$, then*

$$\|\Lambda_V^{-1}(\hat\theta_U-\hat\theta_V)\|\le C_2|\det G_V|^{-1}(\sigma^2N_V^{-1}\log N_U)^{1/2}$$

where $\|\theta\|^2=\theta_0^2+\ldots+\theta_{m-1}^2$ *and*

$$C_2=\left(m+2\sqrt{m}\right)\sqrt{2(\alpha+p)}.$$

Particularly,

$$|\hat{f}_U(x_0) - \hat{f}_V(x_0)| \le C_2 |\det G_V|^{-1} (\sigma^2 N_V^{-1} \log N_U)^{1/2}$$

and

$$|\hat{\theta}_{U,k} - \hat{\theta}_{V,k})| \le C_2 d_{V,2k}^{-1/2} |\det G_V|^{-1} (\sigma^2 N_V^{-1} \log N_U)^{1/2}, \qquad k = 0, 1, \dots, m-1.$$

Proof. Let $\tau_{U,V}$ be m-vector with coordinates

$$\begin{aligned} \tau_{U,V,k} &= \sigma N_V^{-1/2} T_{U,V,k} = \frac{1}{N_V \sqrt{d_{V,2k}}} \sum_V (X_i - x_0)^k \varepsilon_{U,i}, \\ &= \frac{1}{N_V \sqrt{d_{V,2k}}} \sum_V (X_i - x_0)^k \left[Y_i - \sum_{k'=0}^{m-1} \hat{\theta}_{U,k} (X_i - x_0)^{k'} \right], \end{aligned}$$

$k = 0, 1, \dots, m-1$. Using matrix notation, we can rewrite this equality in the form

$$\tau_{U,V} = N_V^{-1} \Lambda_V \left(\Sigma_V Y_V - \Sigma_V \Sigma_V^T \hat{\theta}_U \right).$$

The definition of the least square estimate $\hat{\theta}_V$ implies the equality

$$\Sigma_V Y_V = \Sigma_V \Sigma_V^T \hat{\theta}_V$$

see (2.3). Hence

$$\tau_{U,V} = N_V^{-1} \Lambda_V \Sigma_V \Sigma_V^T \left(\hat{\theta}_V - \hat{\theta}_U \right) = \Lambda_V D_V (\hat{\theta}_V - \hat{\theta}_U).$$

When denoting

$$\eta_{U,V} = \Lambda_V^{-1} (\hat{\theta}_V - \hat{\theta}_U), \tag{3.8}$$

we get

$$\tau_{U,V} = G_V \eta_{U,V}. \tag{3.9}$$

The fact that $\varrho_{U,V} = 0$ means

$$|\tau_{U,V,k}| \le r$$

where

$$r = N_V^{-1/2} \sigma (2 + \sqrt{m}) \sqrt{2(\alpha + p) \log N_U}.$$

Particularly

$$\|\tau_{U,V}\|^2 := \sum_{k=0}^{m-1} \tau_{U,V,k}^2 \le m r^2. \tag{3.10}$$

It remains to understand what follows from this inequality for the vector $\eta_{U,V} = G_V^{-1} \tau_{U,V}$ see (3.9).

Using Cauchy-Schwarz inequality, one gets by (2.8),

$$g_{V,k,k'}^2 = \frac{d_{V,k+k'}^2}{d_{V,2k} d_{V,2k'}} \le 1.$$

Hence the matrix G_V is symmetric with elements at most 1 in absolute value. Therefore all eigenvalues of this matrix are also at most 1 in absolute value. This implies easily that all eigenvalues of G_V^{-1} are at most $|\det G_V|^{-1}$. Therefore, using (3.10)

$$\|\eta_{U,V}\| = \|G_V^{-1}\tau_{U,V}\| \le r\sqrt{m}|\det G_V|^{-1}.$$

In view of (3.8), the assertion follows. □

The next statement is nothing else as the standard decomposition of the local polynomial estimator into deterministic and stochastic terms, compare Cleveland (1979), Katkovnik(1985), Korostelev and Tsybakov (1993), Goldenshluger and Nemirovski (1994). Particularly it shows that if the function f is regular on U and the matrix G_U is well defined, then the estimator $\hat{\theta}_U$ provides a good accuracy of estimation of the function f and its derivatives at x_0.

Proposition 3.3. *Let $U \in \mathcal{U}$ and let G_U be non-singular, see (3.7). Let also*

$$N_U^{-1} \sum_U |f(X_i) - P_\theta(X_i - x_0)|^2 \le \delta_U^2 \tag{3.11}$$

with some $\delta_U > 0$ and $\theta = (\theta_0, \dots, \theta_{m-1})$. Here $P_\theta(z) = \theta_0 + \theta_1 z + \ldots + \theta_{m-1} z^{m-1}$. Then one has for the vector $\hat{\theta}_U$ from (2.4)

$$\Lambda_U^{-1}(\hat{\theta}_U - \theta) = \delta_U G_U^{-1} w_U + \sigma N_U^{-1/2} G_U^{-1/2} \gamma_U \tag{3.12}$$

where $w_U = (w_{U,0}, \dots, w_{U,m-1})$ is a non-random vector in R^m such that

$$|w_{U,k}| \le 1, \qquad k = 0, \dots, m-1, \tag{3.13}$$

$$\gamma_U \sim \mathcal{N}(0, I_m), \tag{3.14}$$

and for every $k = 0, 1, \dots, m-1$

$$\hat{\theta}_{U,k} - \theta_k = d_{U,2k}^{-1/2} |\det G_U|^{-1} (z_1 \delta_U + z_2 \sigma N_U^{-1/2} \gamma'_{U,k}) \tag{3.15}$$

where $|z_1| \le 1$, $|z_2| \le 1$ and $\gamma'_{U,k} \sim \mathcal{N}(0,1)$.

Proof. Denote $\eta_U = \Lambda_U^{-1}(\hat{\theta}_U - \theta)$. Then, using (2.4), (1.1) and (3.7), we obtain

$$\begin{aligned} \eta_U &= \Lambda_U^{-1}(\Sigma_U \Sigma_U^T)^{-1} \Sigma_U (Y_U - \Sigma_U^T \theta) \\ &= N_U^{-1} G_U^{-1} \left[\Lambda_U \Sigma_U (f_U - \Sigma_U^T \theta) + \Lambda_U \Sigma_U \xi_U\right] = \\ &= \delta_U G_U^{-1} w_U + \sigma N_U^{-1/2} G_U^{-1/2} \gamma_U. \end{aligned}$$

Here f_U means the vector in R^{N_U} with elements $f(X_i)$, $X_i \in U$. Also we denoted by w_U a non-random vector in R^m defined by $w_U = \delta_U^{-1} \Lambda_U \Sigma_U (f_U - \Sigma_U^T \theta)$ and by γ_U a random vector in R^m with $\gamma_U = \sigma^{-1} G_U^{-1/2} \Lambda_U \Sigma_U \xi_U$.

For (3.12), it remains to check (3.13) and (3.14). Note that

$$(f_U - \Sigma_U \theta)_i = f(X_i) - \sum_{k=0}^{m-1} \theta_k (X_i - x_0)^k$$

and in view of (3.11)

$$N_U^{-1} \sum_U |(f_U - \Sigma_U \theta)_i|^2 \le \delta_U^2.$$

Next, using the Cauchy-Schwarz inequality

$$\begin{aligned}|w_{U,k}| &= \delta_U^{-1} d_{U,2k}^{-1/2} \left| \sum_U (X_i - x_0)^k (f_U - \Sigma_U \theta)_i \right| \\ &\le \delta_U^{-1} \left[N_U d_{U,2k}^{-1} \sum_U (X_i - x_0)^{2k} \right]^{1/2} \left[N_U^{-1} \sum_U (f_U - \Sigma_U \theta)_i^2 \right]^{1/2} \le 1.\end{aligned}$$

Finally we observe that γ_U is a Gaussian vector with the covariance matrix

$$\mathbf{E}\gamma_U \gamma_U^T = \sigma^{-2} N_U^{-1} G_U^{-1/2} \Lambda_U \Sigma_U \mathbf{E}\xi_U \xi_U^T \Sigma_U^T \Lambda_U G_U^{-1/2} = I_m.$$

The statement (3.15) is a consequence of (3.12). In fact, let us fix some $k \in \{0, 1, \dots, m-1\}$. Then $d_{U,2k}^{1/2}(\hat{\theta}_{U,k} - \theta_k)$ is kth component of $\Lambda_U^{-1}(\hat{\theta}_U - \theta)$. Next, arguing as at the end of the proof of Proposition 3.2 we obtain that $|(G_U^{-1} w_U)_k| \le |\det G_U|^{-1}$. In the similar way, the kth component $\gamma'_{U,k}$ of the Gaussian vector $G_U^{-1/2}\gamma_U$ is a Gaussian random variable with zero mean and $\mathbf{E}(\gamma'_{U,k})^2 \le |\det G_U|^{-1} \le |\det G_U|^{-2}$. This implies (3.15). $\square$

The next result can be viewed as a complement to Proposition 3.1. In contrast, we consider the case when there is a change-point inside the considered interval U. We shall show that if the size of jump in the function itself or in any of its derivatives up to the order $m-1$ is large enough, then the interval U will be rejected with probability close to 1.

Proposition 3.4. *Let $U \in \mathcal{U}$ and let there be $V_1, V_2 \in \mathcal{V}(U)$ such that*

$$N_{V_j}^{-1} \sum_{V_j} |f(X_i) - P_{\theta_{V_j}}(X_i - x_0)|^2 \le \delta_{V_j}^2, \qquad j = 1, 2, \tag{3.16}$$

where $\theta_{V_1}, \theta_{V_2}$ are vectors of coefficients and $\delta_{V_1}, \delta_{V_2}$ are some positive constants. If, for some $k = 0, \dots, m-1$,

$$|\theta_{V_1,k} - \theta_{V_2,k}| \ge b_{V_1,k} + b_{V_2,k} \tag{3.17}$$

with

$$b_{V,k} = d_{V,2k}^{-1/2} |\det G_V|^{-1} \left[C_3 \sigma N_V^{-1/2} \sqrt{\log N_U} + \delta_V \right] \tag{3.18}$$

where V equals V_1 or V_2 and

$$C_3 = C_2 + \sqrt{2p} = \sqrt{2p} + (m + 2\sqrt{m})\sqrt{2(\alpha + p)}, \tag{3.19}$$

then

$$\mathbf{P}_f(\varrho_U = 0) \le N_U^{-p}. \tag{3.20}$$

Proof. The event $\varrho_U = 0$ implies $\varrho_{U,V_j} = 0$, $j = 1, 2$. Let V be V_1 or V_2. By Proposition 3.2

$$|\hat{\theta}_{U,k} - \hat{\theta}_{V,k}| \le C_2 |\det G_V|^{-1} d_{V,2k}^{-1/2} (\sigma^2 N_V^{-1} \log N_U)^{1/2}.$$

Next, by application of Proposition 3.3 one gets

$$\hat{\theta}_{V,k} - \theta_{V,k} = d_{V,2k}^{-1/2} |\det G_V|^{-1} [z_1 \delta_V + z_2 \sigma N_V^{-1/2} \gamma_{V,k}]$$

with δ_V from (3.16), $|z_1|, |z_2| \le 1$ and $\gamma_{V,k} \sim \mathcal{N}(0,1)$. Along with these inequalities and (3.18) we obtain

$$\mathbf{P}_f\left(|\hat{\theta}_{U,k} - \theta_{V,k}| > b_{V,k}\right) \le \mathbf{P}\left(|\gamma_{V,k}| > \sqrt{2p\log N_U}\right) \le N_U^{-p}, \qquad V = V_1 \text{ or } V_2.$$

This and (3.17) obviously imply (3.20). □

Now we are ready to formulate the main results. We distinguish between two different situations. First we assume that the function f has no change-points in the vicinity of the point x_0. In this case, the "ideal" window is to be "symmetric". Then we explore the case with presence of change-points near the point of estimation x_0. For such a situation, the "ideal" window is to be one-side oriented.

To begin by, we introduce the class of "symmetric" windows. Let us fix some positive d_0. We say that some window $U = [x_0 - a_1, x_0 + a_2]$ from $\mathcal{U}$ belongs to the class $\mathcal{U}_s(d_0)$ if, for $V_1 = [x_0 - a_1, x_0]$, $V_2 = [x_0, x_0 + a_2]$, one has

$$\begin{aligned} &1/2 \le N_{V_1}/N_{V_2} \le 2, \\ &|\det G_{V_1}| \ge d_0, \\ &|\det G_{V_2}| \ge d_0. \end{aligned}$$

The first condition here justifies the using of the termin "symmetric window" for each $U \in \mathcal{U}_s(d_0)$.

Theorem 3.1. *Suppose that* $|f(x_0)| \le f_0$. *Let, for some* $d_0 > 0$, *there be a window* $U = [x_0 - a_1, x_0 + a_2]$ *from* $\mathcal{U}_s(d_0)$ *satisfying also (3.1),* $U \in \mathcal{U}^+ \cap \mathcal{U}_s(d_0)$. *Then*

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \le (C_4\sigma^2 N_U^{-1}\log n)^{p/2} + m(2f_0)^p N_U^{-p/2}$$

where

$$C_4 = 3d_0^{-2}\left[2C_1 + C_2 + C(p)\right]^2 = 3d_0^{-2}\left[(m + 2 + 2\sqrt{m})\sqrt{2(\alpha + p)} + C(p)\right]^2, \tag{3.21}$$

and $C(p) \le 2$.

Discussion 3.1. The above result prompts the following definition of the "ideal symmetric" window U_f,

$$U_f = \operatorname{argmax}\{N_U : U \in \mathcal{U}^+ \cap \mathcal{U}_s(d_0)\}.$$

The statement of Theorem 3.1 shows that the adaptive procedure provides with the accuracy of estimation of the same order as if the "optimal" window U_f were known and if we just apply the corresponding estimator $\hat{f}_{U_f}$.

Proof. Let U^* be selected by the adaptive procedure, see (2.9). We distinguish between two cases: $N_{U^*} < N_U$ and $N_{U^*} \ge N_U$. (Recall that due to Proposition 3.1, one has $\varrho_U = 0$ with probability close to 1 and hence typically $N_{U^*} \ge N_U$.)

Note first that, by construction, $|\hat{f}_{x_0}| \le f_0$ and by theorem's condition $|f(x_0)| \le f_0$. Hence $|\hat{f}(x_0) - f(x_0)| \le 2f_0$ and

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \mathbf{1}(N_{U^*} < N_U) \le (2f_0)^p \mathbf{P}_f(N_{U^*} < N_U).$$

Obviously $\mathbf{P}_f(N_{U^*} < N_U) \le \mathbf{P}_f(\varrho_U = 1)$ and by Proposition 3.1 we obtain

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \mathbf{1}(N_{U^*} < N_U) \le (2f_0)^p m N_U^{-p}. \tag{3.22}$$

Next we consider the case with $N_{U^*} \ge N_U$. Obviously U^* contains either $[x_0 - a_1, x_0]$ or $[x_0, x_0 + a_2]$. By making use of the definition of the class $\mathcal{U}_s(d_0)$, we get for $V = U \cap U^*$ that $N_V \ge \min\{N_{V_1}, N_{V_2}\} \ge N_U/3$ and $|\det G_V| \ge d_0$. The fact that $\varrho_{U^*} = 0$ implies in particular that $\varrho_{U^*,V} = 0$. Using now the result of Proposition 3.2 we conclude that

$$|\hat{f}_{U^*}(x_0) - \hat{f}_V(x_0)| \le C_2(\sigma^2 N_V^{-1} \log N_{U^*})^{1/2}. \tag{3.23}$$

Next, since $V \subset U$, then $\Delta_V(f) \le \Delta_U(f)$ and the application of Proposition 3.3 to $\hat{f}_V(x_0)$ gives

$$\hat{f}_V(x_0) - \theta_{V,0} = \sigma N_V^{-1/2} |\det G_V|^{-1} \left[z_{V,1} C_1 \sqrt{\log N_U} + z_{V,2}\gamma_{V,0} \right] \tag{3.24}$$

where $|z_{V,1}|, |z_{V,2}| \le 1$ and $\gamma_{V,0} \sim \mathcal{N}(0,1)$. From the definition of $\Delta_V(f)$ it follows that $|f(x_0) - \theta_{V,0}| \le \Delta_V(f) \le \Delta_U(f)$. Along with (3.23) and (3.24) and applying $|\det G_V|^{-1} \le d_0^{-1}$, we conclude

$$\begin{aligned}
&\mathbf{E}_f|(\hat{f}(x_0) - f(x_0)|^p \mathbf{1}(N_{U^*} \ge N) \\
&\quad \le \mathbf{E}_f \left| \hat{f}_{U^*}(x_0) - \hat{f}_V(x_0) + \hat{f}_V(x_0) - \theta_{V,0} + \theta_{V,0} - f(x_0) \right|^p \\
&\quad \le \sigma^p N_V^{-p/2} d_0^{-p} \mathbf{E}|(2C_1 + C_2)\sqrt{\log n} + \gamma_{V,0}|^p \\
&\quad \le [2C_1 + C_2 + C(p)]^p \sigma^p d_0^{-p} (3N_U^{-1}\log n)^{p/2}.
\end{aligned}$$

Here we have used the inequality $\mathbf{E}|\varkappa + \xi|^p \le (\varkappa + C(p))^p$ for a standard normal ξ and some positive constant $C(p) \le 2$. This and (3.22) prove the assertion. $\square$

Now we are in a position to state the result about the quality of estimation near a change-point. For this we have to be more definitive with our procedure. We will assume that the set $\mathcal{U}$ is defined as above in Section 2 by two sets of endpoints $\mathcal{A}_l$ and $\mathcal{A}_r$,

$$\mathcal{U} = \{U = [a_l, a_r] : a_l \in \mathcal{A}_l, a_r \in \mathcal{A}_r, N_U \ge m\}.$$

Let also $\mathcal{A} = \mathcal{A}_l \cup \mathcal{A}_r$ and let, for each $U \in \mathcal{U}$, the set $\mathcal{V}(U)$ be due to (2.6),

$$\mathcal{V}(U) = \{V = [a_-, a_+] : a_-, a_+ \in \mathcal{A},\, V \subseteq U,\, N_V \ge m\}.$$

Our change-point analysis will be based on Proposition 3.4. Let x_{cp} be the location of a change-point. We suppose that the function f is regular in some local left and right neighborhoods of x_{cp}. Let a_1 and a_2 be the closest from the left and from the right to x_{cp} points of the grid $\mathcal{A}$. We suppose in what follows that there are two intervals V_1 from the left of x_{cp} with the right end-point at a_1 and similarly V_2 from the right of x_{cp} with the left end-point at a_2 and such that (3.17) holds. We denote also by V the interval $[a_1, a_2]$ between V_1 and V_2. The result stated below describes the quality of estimation at a point x_0 which lies beyond V_1, V, V_2. We are therefore interested to take V_1 and V_2 as small as possible but the size should be enough to provide (3.17), for more discussion see the next section.

To be more definitive, let us assume that the point x_0 is from the right of V_2. Our assumption that the function f is regular from the right of x_{cp} can be formally described in a way that there is some $U \in \mathcal{U}$ with the left end-point at a_2 such that $\Delta_U(f)$ is small enough, particularly (3.1) is fulfilled.

Theorem 3.2. *Let the function f be bounded by f_0. Let V_1, V_2 and V be introduced above and let vectors $\theta_{V_1}, \theta_{V_2}$ be such that*

$$N_{V_j}^{-1} \sum_{V_j} |f(X_i) - P_{\theta_{V_j}}(X_i - x_0)|^2 \le \delta_{V_j}^2, \qquad j = 1, 2.$$

and also, for some $d_0 > 0$,

$$|\det G_{V_j}| \ge d_0, \qquad j = 1, 2.$$

Next, let for some $k = 0, 1, \ldots, m-1$,

$$|\theta_{V_1,k} - \theta_{V_2,k}| \ge b_{V_1,k} + b_{V_2,k}$$

where, for V equal to V_1 or V_2,

$$b_{V,k} = d_{V,2k}^{-1/2} |\det G_V|^{-1} \left[C_3 \sigma N_V^{-1/2} \sqrt{\log n} + \delta_V \right]$$

C_3 being from (3.19).

Let also U be some interval from $\mathcal{U}^+$, see (3.2), with the left end-point a_2 and such that $V_2 \subset U$, $x_0 \in U \backslash V_2$ and

$$N_{V_1} + N_V + N_{V_2} \le \beta N_U \tag{3.25}$$

with some $\beta < 1$. Then

$$\mathbf{E}|\hat{f}(x_0) - f(x_0)|^p \le \left[(1-\beta)^{-1} C_4 \sigma^2 N_U^{-1} \log N_U \right]^{p/2} + (m+1)(2f_0)^p N_U^{-p/2}$$

where C_4 is as in Theorem 3.1.

Proof. By Proposition 3.1,

$$\mathbf{P}(\varrho_U = 1) \le m N_U^{-p}$$

and by Proposition 3.4, if some U' contains V_1 and V_2 and if $N_{U'} \ge N_U$, then

$$\mathbf{P}(\varrho_{U'} = 0) \le N_U^{-p}.$$

Using the arguments from the proof of Theorem 3.1 we can reduce our consideration to the case when $\varrho_U = 0$ and $\varrho_{U'} = 1$ for every U' with $V_1 \cup V_2 \subset U'$.

Let U^* be selected by the adaptive procedure. Since $\varrho_U = 0$, the definition of U^* implies $N_{U^*} \ge N_U$. Next, U^* does not contain V_1. (Otherwise, U^* contains also V_2 because $x_0 \in U^*$ and V_2 is between V_1 and x_0, hence $\varrho_{U^*} = 1$ does hold.) Denote $U_1 = U \cap U^*$. Condition (3.25) implies that

$$N_{U_1} \ge (1-\beta) N_U. \tag{3.26}$$

In fact, if $U \subseteq U^*$, then $N_{U_1} = N_U$. If $U \not\subseteq U^*$, since also $V_1 \not\subseteq U^*$, then obviously

$$N_{U^*} \le N_{V_1} + N_V + N_{U_1}$$

and (3.26) follows from (3.25). By similar reason, U_1 contains V_2 and $|\det G_{U_1}| \ge |\det V_2| \ge d_0$.

Now, by Proposition 3.2,

$$|\hat{f}_{U^*}(x_0) - \hat{f}_{U_1}(x_0)| \le C_2 |\det G_{U_1}|^{-1} (\sigma^2 N_{U_1}^{-1} \log n)^{-1/2}$$

and by Proposition 3.3,

$$\hat{f}_{U_1}(x_0) - f(x_0) = \sigma N_{U_1}^{-1/2} |\det G_{U_1}|^{-1} [z_1 C_1 \sqrt{\log N_{U_1}} + z_2 \gamma]$$

where $|z_1|, |z_2| \le 1$ and $\gamma \sim \mathcal{N}(0,1)$.

These inequality allow to complete the proof in the same way as for Theorem 3.1.

□

4. The case of an equidistant design

Below we specify the above general results to the case of an equidistant design with the aim to compare our results with the existing in the literature.

We consider the regression model (1.1) with n the design points $X_i = i/n$ in the interval $[0,1]$. We will examine our procedure with the "maximal" $\mathcal{U}$ from (2.1).

First we notice that for every interval U with $N_U \ge m$, one has

$$|\det G_U| \ge d_0$$

where $d_0 > 0$ depends only on m.

We begin by reformulating the statement of Theorem 3.1 for windows U of the form $U = [x_0 - h, x_0 + h]$ with $h = k/n$, $k = m, m+1, \ldots, n$. Obviously $N_U \ge nh + 1$ and $N_U = 2nh + 1$ if $U \in [0,1]$.

Theorem 4.1. *Let $|f(x_0)| \le 1$ and let h be such that for $U = [x_0 - h, x_0 + h] \cap [0,1]$,*

$$\Delta_U(f) \le C_1 \sigma (h^{-1} n^{-1} \log n)^{1/2}, \tag{4.1}$$

where $C_1 = \sqrt{2(\alpha + p)}$, see Theorem 3.1. Then

$$\mathbf{E}_f |\hat{f}(x_0) - f(x_0)|^p \le 2(C_4 \sigma^2 h^{-1} n^{-1} \log n)^{p/2}$$

where C_4 is due to (3.21).

Discussion 4.1. Now we can also reformulate the definition of the "ideal symmetric window" U_f (see the discussion after Theorem 3.1) in terms of "ideal bandwidth" h_f:

$$h_f = \operatorname{argmax}\{h : \Delta_{[x_0-h, x_0+h]}(f) \le C_1 \sigma (h^{-1} n^{-1} \log n)^{1/2}\}. \tag{4.2}$$

The statement of Theorem 4.1 shows that the adaptive procedure provides with the accuracy of estimation corresponding to the choice of the "ideal bandwidth" h_f.

It was proved in Lepski, Mammen and Spokoiny (1994) that each estimation procedure with such properties is automatically rate-optimal for a wide range of Sobolev or Besov classes.

Note that more standard way to define the "ideal bandwidth" is based on the assumption that the function f is m times differentiable and the mth derivative $f^{(m)}$ is uniformly bounded (at least in some neighborhood of the point x_0),

$$|f^{(m)}(x)| \le M m!.$$

In this case one has easily $\Delta_{[x_0-h, x_0+h]} \le M h^m$ and the balance equation $M h^m \approx \sigma h^{-1} n^{-1} \log n$ leads to the bandwidth $h_f \approx (\sigma^2 M^{-2} n^{-1} \log n)^{1/(2m+1)}$. However, our smoothness condition (4.1) is weaker than the usual one and hence the balance rule (4.2) seems to be more flexibile than the last one.

Now we turn to the case when change-points are incorporated in the model. Let x_{cp} be a change-point. Without loss of generality we may assume that x_{cp} coincides with a grid point $a_i = i/n$. As above in Theorem 3.2 we assume that the function f is regular from the left and from the right of x_{cp} and it has a jump of kth derivative at x_{cp} with k from 0 to $m-1$. This is understood in the following way. Let some small $h_0 > 0$ be fixed and let

$$V_1 = [x_{\mathrm{cp}} - h_0, x_{\mathrm{cp}}),$$
$$V_2 = (x_{\mathrm{cp}}, x_{\mathrm{cp}} + h_0].$$

Let also θ_{V_1} and θ_{V_2} be the coefficients of the approximating polynomials for V_1 and V_2. A jump of kth derivative of f means that $\theta_{V_1,j}$ and $\theta_{V_2,j}$ are equal or very close to each other for $j = 0, \dots, k-1$ and the difference $\theta_{V_1,k} - \theta_{V_2,k}$ differs significantly from zero.

We are mostly interested to describe the minimal distance h_0 between the change-point x_{cp} and the point of estimation x_0 which is enough for rate-consistent estimation of $f(x_0)$. Particularly, it is of interest to understand how this distance h_0 depends on what derivative $f^{(k)}$ has a jump and on the jump size.

Theorem 4.2. *Let the function f be bounded by 1. Let h_0, V_1, V_2, θ_{V_1} and θ_{V_2} be introduced above and let, for some k from 0 to $m-1$, one has*

$$|\theta_{V_1,k} - \theta_{V_2,k}| \ge 2b.$$

Let also there be some $h > 2h_0$ such that

$$\Delta_{(x_0,x_0+h]}(f) \le C_1\sigma(h^{-1}n^{-1}\log n)^{1/2}, \qquad (4.3)$$
$$\Delta_{[x_0-h,x_0)}(f) \le C_1\sigma(h^{-1}n^{-1}\log n)^{1/2}$$

with C_1 from Proposition 3.1. If

$$h_0^{2k+1} \ge C_5 b^{-2}\sigma^2 n^{-1}\log n$$

with

$$C_5 = (C_3 + C_1)^2 d_0^{-2}(2k+1) = (2k+1)\left[\sqrt{2p} + (m+1+2\sqrt{m})\sqrt{2(\alpha+p)}\right]^2 d_0^{-2}$$

then for each $x_0 \in [x_{\mathrm{cp}} + h_0, x_{\mathrm{cp}} + h]$ or $x_0 \in [x_{\mathrm{cp}} - h, x_{\mathrm{cp}} - h_0]$, one has

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \le 2(2C_4\sigma^2 h^{-1}n^{-1}\log n)^{p/2}$$

where C_4 is from Theorem 3.2.

Proof. We derive this result from the general result of Theorem 3.2. First we assume without loss of generality that

$$N_{V_1} = N_{V_2} = nh_0$$

and similarly for $U = (x_0, x_0 + h]$

$$N_U = nh.$$

Now condition (4.3) means that $U \in \mathcal{U}^{+}$, see (3.2), and condition (3.25) of Theorem 3.2 is fulfilled with $\beta = 1/2$. Next, one has easily for $V = V_1$ or $V = V_2$ and $x_0 \ge x_{\mathrm{cp}} + h_0$

$$d_{V,2k} = (nh_0)^{-1} \sum_{X_i \in V} (X_i - x_0)^{2k} \ge h_0^{2k}/(2k+1).$$

Therefore, all the conditions of Theorem 3.2 are satisfied and the application of this theorem leads to the desirable assertion. □

Discussion 4.2. We see from the above result that the presence of a change-point leads to poor quality of estimation only in some neighborhood of this change-point. The radius h_0 of this neighborhood depends on the type of change (jump of a function itself or its k th derivative) and on the size b of jump,

$$h_0 \asymp \left(b^{-2} n^{-1} \log n\right)^{1/(2k+1)}.$$

Therefore, the proposed estimation procedure is able to detect about $b^2 n/\log n$ (in order) jumps of a size $b > 0$. Similarly, for jumps of k th derivatives, the detectable number of change-points is about $(b^2 n/\log n)^{1/(2k+1)}$.

At the conclusion, we discuss shortly the question of optimal change-point estimation. It is well known that a single jump can be estimated with the rate n^{-1}, see, for example, Hinkley (1970), Ibragimov and Khasminski (1981), Korostelev (1987). Our procedure provides with the rate $n^{-1}\log n$. The following result shows that this extra log-factor is not only the price for adaptation. Even in the case when at most two jumps are allowed, their locations can be estimated only with the rate $n^{-1}\log n$. Similarly one can show that the optimal rate for estimation of a jump of k th derivative is $(n^{-1}\log n)^{1/(2k+1)}$, if more than one jump is considered.

Introduce the class $\mathcal{F}_h$ of functions with two values $0, 1$ and having two jumps at points x_1 and x_2 inside the interval $[0,1]$ separated with the distance h,

$$|x_1 - x_2| \ge h.$$

Theorem 4.3. *There exists $C > 0$ such that for $h(n) = Cn^{-1}\log n$ and for arbitrary estimates $\hat{x}_1, \hat{x}_2$, the following asymptotic bound holds*

$$\sup_{f \in \mathcal{F}_{h(n)}} \max\{\mathbf{P}_f(|\hat{x}_1 - x_1| > h(n)), \mathbf{P}_f(|\hat{x}_2 - x_2| > h(n))\} \to 1, \qquad n \to \infty.$$

Proof. As usual for such kind of results, we change the minimax problem by a specific Bayes one. Let some positive $C < 2$ be fixed. Set $h(n) = Cn^{-1}\log n$. Without loss of generality we assume that $nh(n) = C\log n$ is an integer number and that $M = 1/h(n) = n/(C\log n)$ is also integer.

Let us split the whole interval $[0,1]$ into M subintervals of length $h(n)$ and denote this partition by $\mathcal{I}$. Each interval I from $\mathcal{I}$ contains $N = nh(n) + 1 = C\log n + 1$ design points.

Now we assume that our function f is random and with probability M^{-1} it coincides with the function f_I which is one on I and zero outside. Now our original problem can be clearly reduced to the problem of estimating I by observed data.

Denote by $Z_{I,n}$ the log-likelihood

$$Z_{I,n} = \log(d\mathbf{P}_{f_I}/d\mathbf{P}_0)$$

where $\mathbf{P}_0$ corresponds to the function $f \equiv 0$. One obtains easily from (1.1) that

$$Z_I = \frac{1}{2} \sum_I \left[Y_i^2 - (Y_i - 1)^2 \right] = \sum_i Y_i - N/2.$$

Now the Bayes estimate $\hat{I}$ of I for the indicator loss function $\mathbf{1}(\hat{I} \neq I)$ is of obvious structure:

$$\hat{I} = \operatorname*{arginf}_I \frac{1}{M} \sum_{I' \neq I} \exp\{Z_{I'}\} = \operatorname*{argmax}_I Z_I.$$

Let us fix an arbitrary $I_0 \in \mathcal{I}$ and consider the probability $\mathbf{P}_{I_0}(\hat{I} \neq I_0)$ where the measure $\mathbf{P}_{I_0}$ corresponds to the function f_{I_0}. First we note that one has under $\mathbf{P}_{I_0}$ with probability 1

$$\sum_{I_0} Y_i = \sqrt{N} \zeta_{I_0} + N,$$
$$\sum_I Y_i = \sqrt{N} \zeta_I, \qquad I \neq I_0$$

where $\zeta_I = N^{-1/2} \sum_I \xi_i$, and obviously all ζ_I are standard normal. Now

$$\mathbf{P}_{I_0}(\hat{I} \neq I_0) = \mathbf{P}\left(\max_{I \neq I_0} \zeta_I - \sqrt{N}/2 > \zeta_{I_0} + \sqrt{N}/2 \right) = \mathbf{P}\left(\max_{I \in \mathcal{I}} \zeta_I > \sqrt{N} \right).$$

It is well known, see e.g. Petrov (1975), that for each $\alpha < 2$

$$\mathbf{P}\left(\max_{I \in \mathcal{I}} \zeta_I > \sqrt{\alpha \log M} \right) \to 1, \qquad M \to \infty.$$

Therefore, the desirable assertion follows if $\alpha \log M > N$ or equivalently

$$C \log n + 1 < \alpha \log(n/(C \log n)).$$

It remains to observe that the latter property holds true for $C < \alpha < 2$ and n large enough. $\square$

References

[1] Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. of the American Stat. Spc.*, **74**, 829–836.

[2] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1994). Wavelet shrinkage: asymptopia? *J. Royal Statist. Soc.*, Ser.B, **57**, 301–369.

[3] Goldenshluger, A. and Nemirovski, A. (1994) On spatial adaptive estimation of nonparametric regression. Technical Report 5/94, Technion, Haifa.

[4] Hall, P., Kerkyacharian, J. and Picard, D. (1996) On the minimax optimality of block thresholded wavelet estimators. Unpublished manuscript.

[5] Hall, P. and Patil, P. (1995) Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23**, 905–928.

[6] Hinkley, D. (1970) Inference about a change point in a sequence of random variables. *Biometrica* **57**, 41–58.

[7] Ibragimov, I. and Hasminski, R. (1981) *Statistical Estimation: Asymptotic Theory.* Springer Verlag, New York–Heidelberg–Berlin.

[8] Katkovnik, Yu.A. (1985) *Nonparametric Identification and Data Smoothing: Local Approximation Approach.* Nauka, Moscow (in Russian).

[9] Korostelev, A. (1987) On minimax estimation of a discontinuous signal. *Theory Probab. Appl.* **32**, 727–730.

[10] Korostelev, A. and Tsybakov, A. (1993) *Minimax Theory of Image Reconstruction.* Springer Verlag, New York–Heidelberg–Berlin.

[11] Lepski, O., Mammen, E. and Spokoiny, V. (1995) Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Annals of Statistics*, to appear.

[12] Lepski, O. and Spokoiny, V. (1994) Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics*, to appear.

[13] Müller, H. (1992) Change-points in nonparametric regression analysis. *Ann. Statist.* **20**, 737–761.

[14] Oudshoorn, C. (1995). Minimax estimation of a regression function with jumps: attaining the optimal constant. Technical Report 934, Department of Mathematics, University Utrecht.

[15] Petrov, V.V. (1975) *Sums of Independent Random Variables.* Springer, New York.

[16] Wang, Y. (1995) Jump and sharp cusp detection by wavelets. *Biometrica* **82**, 385–397.

[17] Wu, J. and Chu, C. (1993) Kernel type estimators of jump points and values of a regression function. *Ann. Statist.* **21**, 1545–1566.

[18] Yin, Y. (1988) Detection of the number, locations and magnitudes of jumps. *Comm. Statist. Stochastic Models* **4**, 445–455.

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany., E-mail: spokoiny@wias-berlin.de